Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/12/eaax7798/DC1

Supplementary Materials for

Distinguishing cell phenotype using cell epigenotype

Thomas P. Wytock* and Adilson E. Motter

*Corresponding author. Email: t-wytock@northwestern.edu

Published 18 March 2020, *Sci. Adv.* **6**, eaax7798 (2020)
DOI: 10.1126/sciadv.aax7798

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/12/eaax7798/DC1)

Supplementary Information

Method testing for synthetic data.

We generated synthetic data as described in Methods under three scenarios to test the efficacy of the KNN eigengene method versus the gene-based method. The first scenario, reflected in fig. S2 (A and B), shows that the correlation-based method does indeed perform better than the gene-based method. In fig. S2A, the rate at which the correlation method correctly identifies at least one of the cell-type defining eigengenes climbs steeply between as the SNR varies from 1 to 10. In fig. S2B, the correlation method, but not the gene method, accurately identifies cell types over this SNR range, as measured by the root mean squared error between the KNN-predicted probability of cell type membership ($\hat{w}_{im}$) and the actual cell type measurement. If the probability of belonging to a cell type were uniformly distributed between 0 and 1, this measurement would converge to $\sqrt{2}$, while perfect prediction corresponds to 0. The deviation from $\sqrt{2}$ for small SNR is attributable to the feature optimization step, as it selects the best feature out of a set of 100. We note that incorrect identification of one of the eigengenes tends to penalize model accuracy. This reflects eigengenes with small eigenvalues being selected to define the difference between cell types, resulting in the effective SNR being much smaller than the nominal SNR. fig. S2 (C and D) shows that the correlation method still manages to perform well, even when the genes define the cell type in lieu of the eigengenes. While the probability of identifying the correct gene increases faster than the probability of identifying the correct eigengene, the gene-based method fails more drastically when it cannot identify the correct gene. The superior performance of the correlation method in this case is explained by the fact that the difference in a given gene is distributed across all of the correlation eigenvectors so that the method is not sensitive to whether the correct gene is deduced or not. Finally, we consider the case in which cell type differences are defined by a correlation difference, but a single gene is spuriously differentially expressed in the training set, but not in the test set (fig. S2 E and F). When the correlation eigenvector can be identified, which happens almost as often as in (fig. S2A), the correlation method uniformly outperforms the gene-based method. Thus, the correlation-based method is robust to single-gene errors and can work even in cases where the cell types are defined by genes. We note that our method performs well because there is an underlying correlation structure to the data in all cases, which is a well grounded assumption for biological systems. In contexts where the underlying variables are *uncorrelated*, we would expect the performance of the correlation-based method to deteriorate relative to single-feature methods.

Assessing unsupervised methods.

Since PDM *(22)* and SC3 *(23)* are unsupervised methods, the number of clusters C that it produces is not constrained to be equal to the number of cell types K. If $C \ll K$, (as is the case in which PDM is applied to the GeneExp dataset), then PDM will necessarily have limited accuracy, but we can assign a cell type to each cluster by determining the cell type of the largest fraction of measurements belonging to that cluster. However, in the case that $C = M$, where M is the number of experiments, assignment of each cluster to the cell type of that experiment would rate as "perfect" prediction. Such a partition is uninformative. Thus, simply assigning each cluster to the largest fraction cell type will overstate the method's accuracy when the clustering method subdivides the experiments to many groups.

Therefore, we calculate the accuracy using the following thought experiment. Suppose that we sample one experiment from each cluster, chosen at random, and determine the cell type. Then $p_i^{(k)} = m_i^{(k)}/m^{(k)}$ is the probability of sampling cell type i in cluster k, where $m^{(k)}$ is the total number of measurements in cluster k, of which $m_i^{(k)}$ belong to cell type i. The average number of experiments correctly predicted in the cluster is

$$n^{(k)} = \sum_{i \in \{k\}} p_i^{(k)} m_i^{(k)}, \tag{1}$$

where {k} is the set of cell types in cluster k. The total fraction predicted is then

$$h = \frac{1}{M} \sum_{k \in C} n^{(k)}. \tag{2}$$

Suppose further that each cell type is only able to be assigned once, then Eq. (1) remains the same, but Eq. (2) must be modified so that no cell type is assigned to two different clusters. We look for the best assignment of cell types to clusters by randomly assigning cell types to each cluster, and continue until either all clusters have a cell type or all cell types have been assigned. We repeat this 1,000 times and take the maximum value found as the method accuracy.

On the role of long-range contacts in Hi-C data.

Previous work has used Hi-C data to investigate the short-range structure ($<$ 500 kb) of chromatin and understand how proteins like CTCF package DNA into loops called TADs *(27)*. In fig. S8, we show that long-range contacts, rather than short-range contacts, arise as important for predicting cell type. This is significant because short-range contacts have been well-studied and are thought to be highly conserved between cell types and even species, whereas less is known about the nature of long-range contacts. To assess the contribution of long-range versus short-range contacts to predict cell type, we removed all contacts in a range either below (fig. S8A) or above (fig. S8B) the distance indicated by the legend. We observe that removal of all contacts < 500 kb does not meaningfully impact the predictive accuracy, but removing contacts above this range causes accuracy to decrease. In addition, when keeping only the local contacts, the method is relatively poor at distinguishing cell types. Keeping contacts in the 500–1000 kb range and the 2.5–10 Mb range appears to enhance predictive accuracy.

To further substantiate whether Hi-C structures are different in different cell types, we employed a functional attribution method in which we removed the contacts of all loci associated with Variable-Diversity-Joining (VDJ) recombination (igH, igK, and igL) *(43)* and re-ran our prediction model. Since VDJ is present in only the B cells in our dataset, the masking of this data should reduce the confidence of classifying B cells, which is exactly what happens. We calculated the number of B cell measurements whose prediction accuracy improves upon inclusion of VDJ loci and found that 40 instances do under the correlation-based model with three eigenloci. We calculated a bootstrapped distribution by randomly selecting the same number of loci and examining how many total B cell measurements improved, and we found that the observed number is > 5 standard deviations larger than the null expectation. This analysis demonstrates that (i) aspects of chromatin structure are cell-type and species specific, as these chromosomal regions are not conserved across organisms or human cell types and (ii) functional attribution is achievable by masking the loci and observing the change in probability.

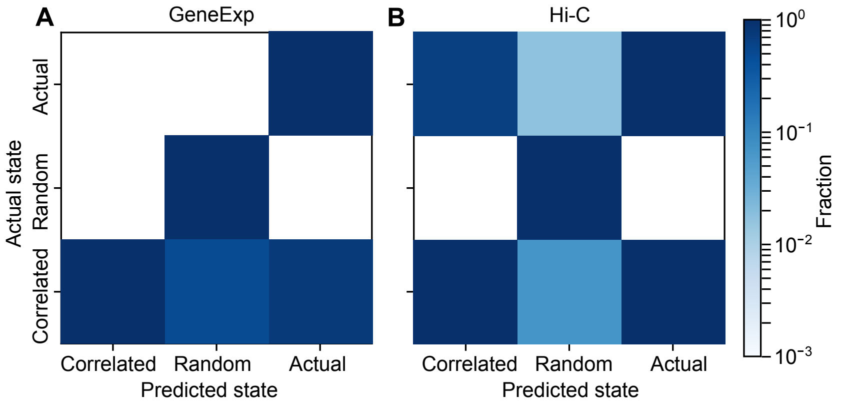

Fig. S1. Confusion matrices for discerning actual and simulated data. (**A**) Distinguishability of actual data from the GeneExp dataset from uncorrelated simulated data (Uncorrelated), and correlated simulated data (Correlated), with accuracies color-coded as a fraction of the number of states predicted. In the confusion matrix, rows correspond to the actual method used to generate the data and columns map to the predicted method used to generate the data. (**B**) Same as (A), but for the Hi-C dataset. In both datasets, the actual data are confused with the simulated, correlated data much more frequently than they are with simulated, uncorrelated data. The misclassification rates are $> 33\%$ for GeneExp and $> 70\%$ for Hi-C.

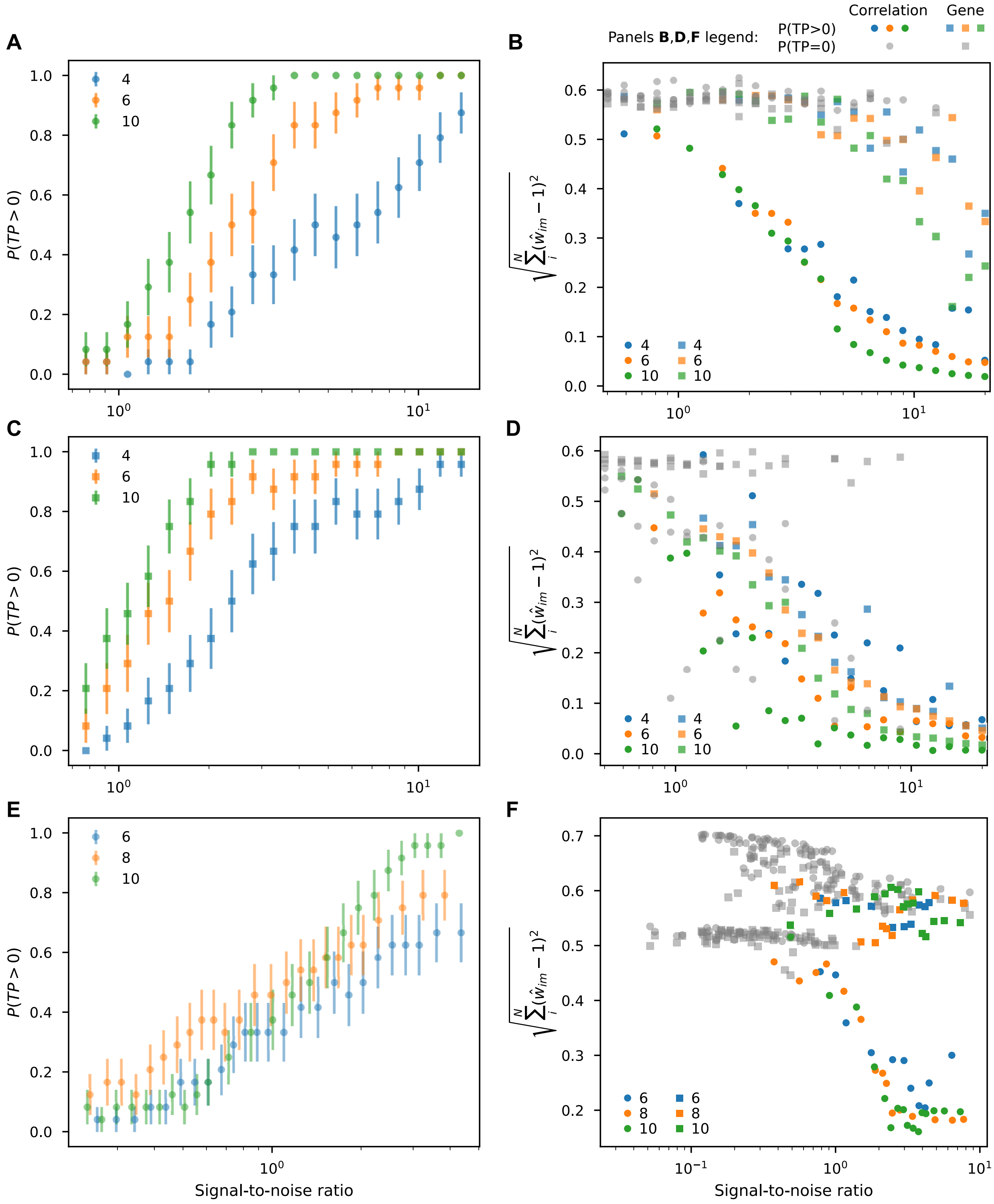

Fig. S2. Method testing results as a function of the SNR under three scenarios (rows) for two criteria (columns). (**A**) Probability of identifying a differentially expressed eigengene as a function of the SNR and the number of experiments (color-coded), with error bars denoting the standard error of the mean, for eigengene-based cell types. (**B**) Root mean square deviation between the KNN-inferred probability $\hat{w}_{im}$ and the actual cell type as a function of SNR. Instances in which differentially expressed genes or eigengenes are not identified are colored in gray. (**C** and **D**) Results for gene-based cell types. Axes, colors, and symbols are as defined in (A and B), respectively. (**E** and **F**) Results for eigengene-based cell types with one confounding differentially expressed gene between the cell types in the training set. Axes, colors, and symbols are as defined in (A and B), respectively.

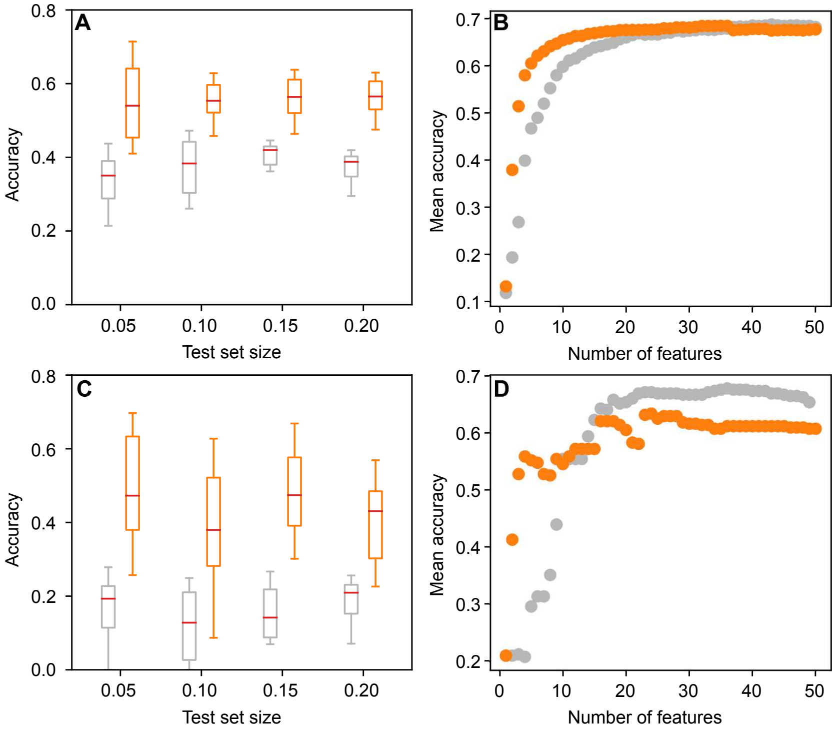

Fig. S3. Comparison of forward selection with PCA. (**A**) Accuracy of PCA (grey) and forward selection (orange) as a function of the test set size, expressed as a fraction of the total number of experiments in the GeneExp dataset. (**B**) Accuracy as a function of the number of features for the GeneExp dataset. (**C**) Same as (A), but for the Hi-C dataset. (**D**) Same as (B), but for the Hi-C dataset. Axes labels for (A–D) retain their meanings from Fig. 4. Differences in all distributions in (A and C) are significant at the $p < 0.01$ level (Kolmogorov-Smirnov test).

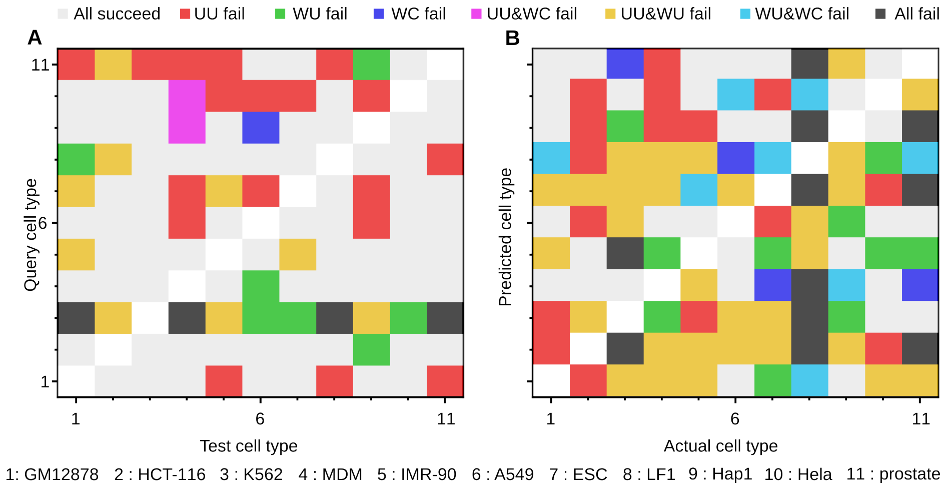

Fig. S4. Distinguishing cell types for the Hi-C dataset. (**A**) Cell type homogeneity, where axes and color code retain their meanings from Fig. 3. (**B**) Nonconvex fraction for a sampling of chords between pairs of same cell type measurements with predicted cell types on the y axis and actual cell types on the x axis. Each square is colored if $> 0.1\%$ of chords of the actual cell type are classified as the predicted cell type using the versions of the method indicated in the legend.

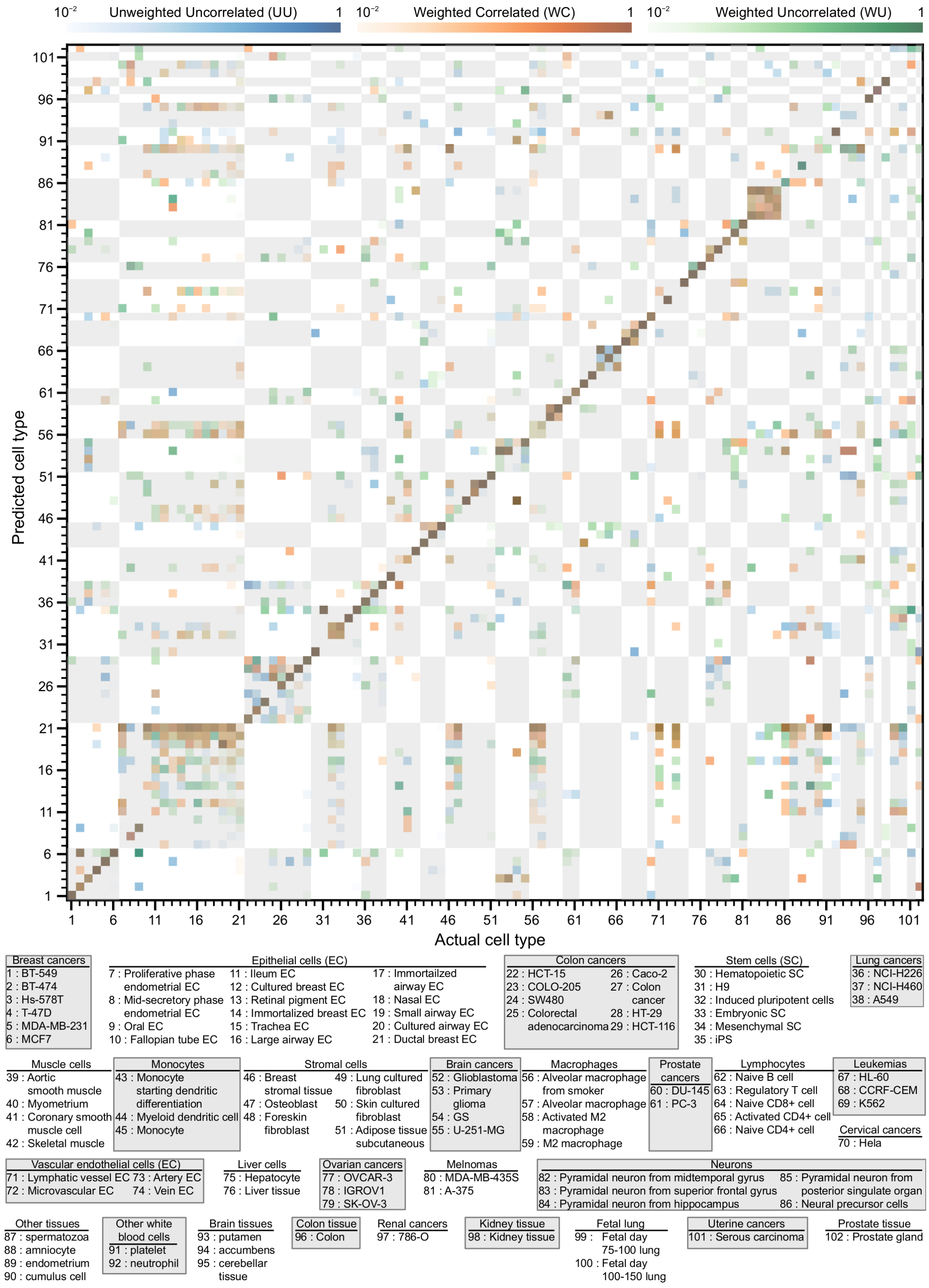

Fig. S5. KNN classification accuracy by cell type for the GeneExp dataset under LOGO cross-validation. The version abbreviations and colors bars are defined in Fig. 5A, and the grey and white checkered background and tick label legend retain their meaning from Fig. 3. The number of experiments for each cell type are listed in table S2. The accuracies averaged by cell type group correspond to those presented in Fig. 5A.

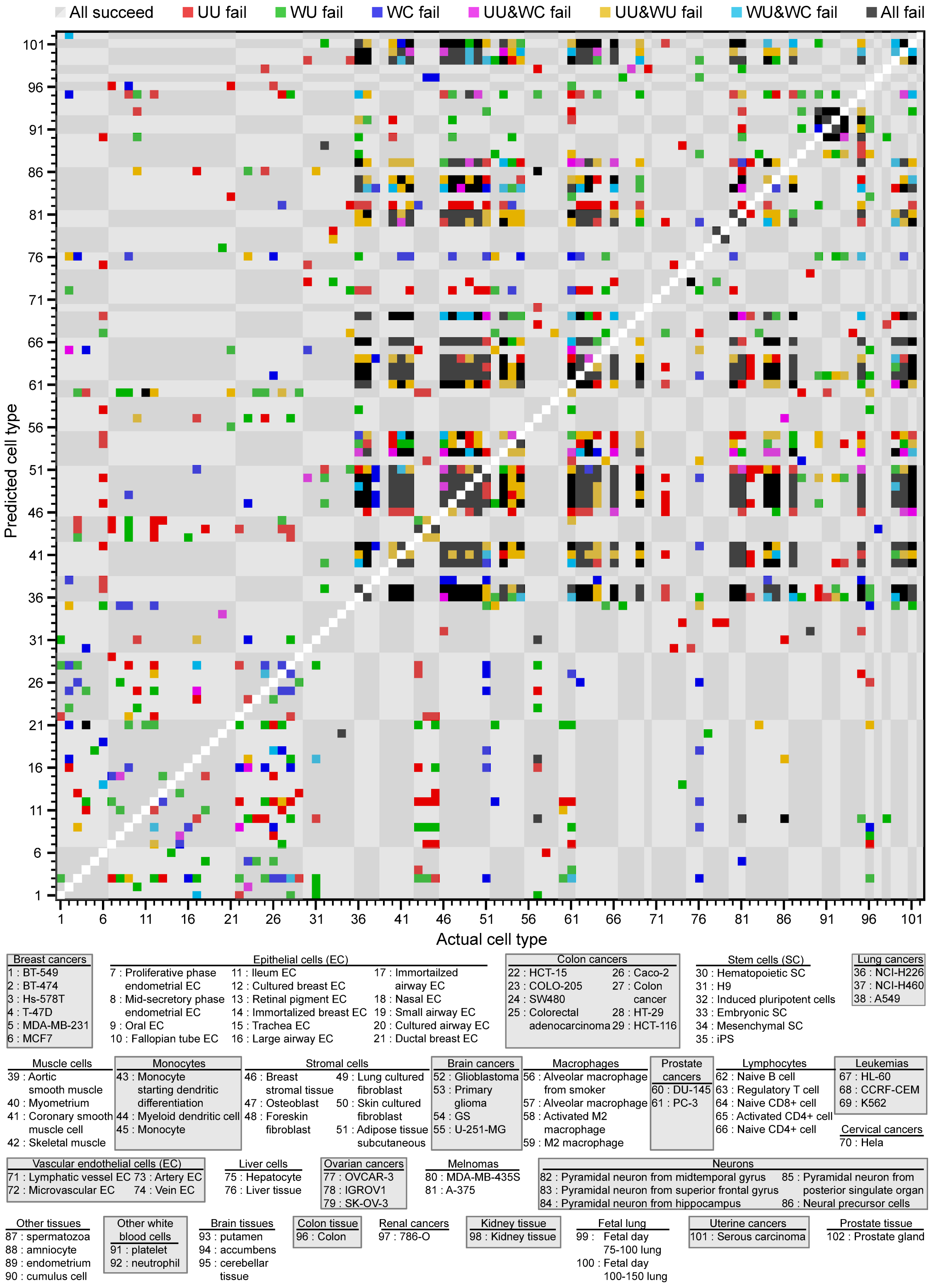

Fig. S6. Fraction of nonconvex chords for each cell type. Colors, background, abbreviations and tick label legend retain their values described in Fig. 3. The predicted cell type was distinguishable from the actual one if $< 0.1\%$ of the chords from the actual cell type (column) were classified as the predicted cell type (row).

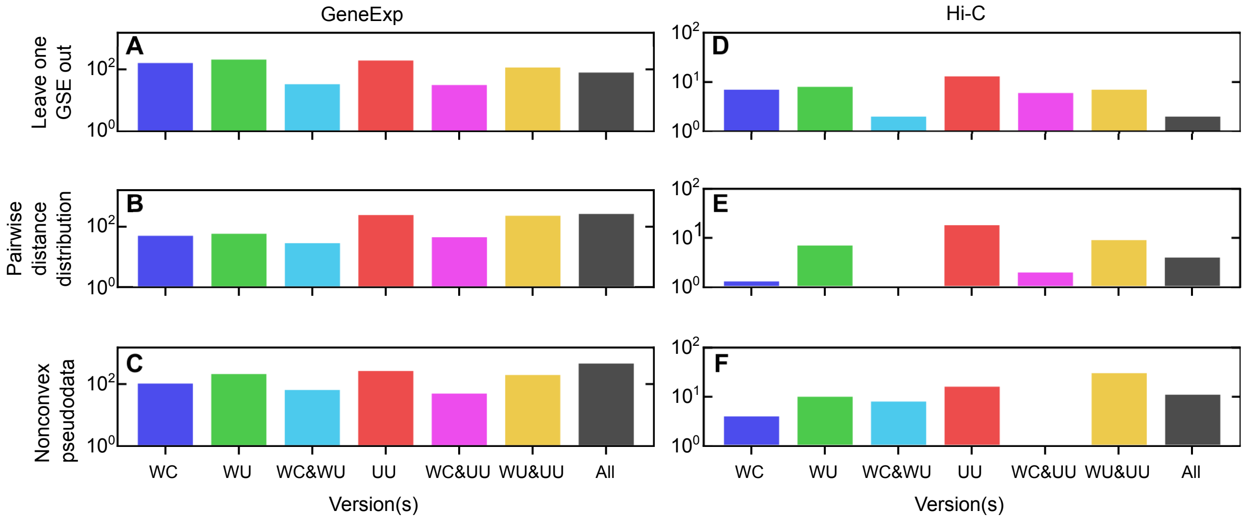

Fig. S7. Compilation of the number of squares of each color found in the preceding figures. (**A**) fig. S5. (**B**) Fig. 3. (**C**) fig. S6. (**D**) Fig. 5B. (**E**) fig. S4A. (**F**) fig. S4B. In (D), cell types are distinguishable (and therefore not counted) if $< 10\%$ of the experiments of a given cell type are classified as that specified by the y axis. The fraction of cases in which all versions of our approach distinguish cell types are not shown to emphasize the differences between the versions.

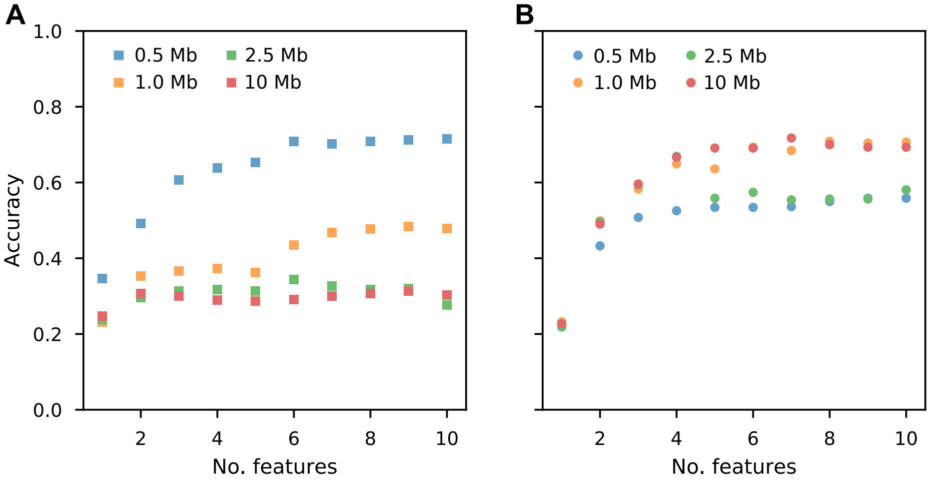

Fig. S8. Accuracy as a function of genomic distance between loci and number of features for the Hi-C dataset. (**A**) Accuracy when classifying on loci pairs separated by more than the distance in megabases (Mb) indicated by the legend. (**B**) Same as (A), but for loci pairs separated by less than the stated distance.

Supplementary Tables

Table S1. Version comparison results and KS test *P* values. Supplementary File: Supplementary Table S1.xlsx

Table S2. Cell type counts, tick labels for Figs. 2C, 3, and 5 and figs. S5 and S6, and database accession numbers for the GeneExp and Hi-C datasets. Supplementary File: Supplementary Table S2.xlsx